MACHINE LEARNING

(CSE 4020)

NAME - CHANDAN CHADHA

REGISTRATION NUMBER - 19BCE1004

SLOT – G2

FACULTY – PROF. SYED IBRAHIM S.P

J COMPONENT

TOPIC: EARLY STAGE MALWARE PREDICTION USING MACHINE LEARNING TECHNIQUES AND RNN

DATASET DETAILS AND DESCRIPTION

For the implementation of the various techniques to test a subject file as malicious or legitimate, I have made the use of the following 2 datasets. The 2 datasets mentioned have been trained and tested using various supervised learning models in Machine Learning and have also been tried using Recurrent Neural Networks that has been built from scratch.

1. CICIDS 2017 DATASET(RECOMMENDED BY THE FACULTY):

Many Intrusion Detection Systems (IDS) has been proposed in the current decade. To evaluate the effectiveness of the IDS Canadian Institute of Cyber security presented a state of art dataset named CICIDS2017, consisting of latest threats and features. The dataset draws attention of many researchers as it represents threats which were not addressed by the older datasets. While undertaking an experimental research on CICIDS 2017, it has been found that the dataset has few major shortcomings. These issues are sufficient enough to bias the detection engine of any typical IDS. This paper explores the detailed characteristics of CICIDS2017 dataset and outlines issues inherent to it. Finally, it also presents a combined dataset by eliminating such issues for better classification and detection of any future intrusion detection engine.

Link for the above mentioned dataset (cicids2017.csv):

https://github.com/datanduth/cicids2017-ml

2. DATA.CSV(CHOSEN BY THE USER):

Dataset downloaded from the Kaggle repository containing details about the data with attributes such as name, size of optional header, characteristics, major linker version, minor linker version, size of code, size of initialized data, size of uninitialized data, base of code, base of data, image base etc and the final class label that assigns the class name to a particular file as malicious or legitimate.

Link for the above mentioned dataset (data.csv):

https://drive.google.com/file/d/1--vgYsVO-b2q90rE6gkq_bL6Cbo2s6D_/view?usp=sharing

INDIVIDUAL CONTRIBUTION

Since, this is project consists of only 1 member (19BCE1004), I have done everything needed for the successful implementation and execution of this project.

MAIN POINTS (in terms of contribution):

- **Data Collection/ Gathering**
- **Data Analysis**
- **Data Pre processing**
- **Choosing the right dataset along with the dataset provided for reference (CICIDS 2017)**
- **Implementation of various Machine Learning Algorithms including Random Forest, Gradient Boosting, GNB (Gaussian Naïve Bayes), Decision Trees (algorithm with the highest accuracy), Linear Regression, Ada Boost**
- **Implementation of Recurrent Neural Networks (RNN) for the same problem statement, building the code and logic of RNN from scratch**
- **Implementation of LSTM and GRU(Gated Recurrent Unit) for the input layer, middle layer(hidden), output layer, building the code and logic of LSTM and GRU from scratch**

CONTRIBUTION TO THE RESEARCH PAPER

- **In order to keep the novelty and uniqueness of the project topic and its implementation through the research paper, I have worked in the R programming environment for analysis of the dataset.**
- **Jupyter Notebook supports the configuration of programming through R Kernel and import the same into an existing Python environment.**
- **For the given dataset given as reference, I have analyzed the dataset and simplified its statistics for all the attributes present in the dataset.**
- **I have read and imported the dataset from the .csv file and print the summary of all its attributes.**
- **I have printed the Confusion Matrix of all the attributes present in the dataset using RNN and Regression Trees.**
- **Based upon the RNN's test and train data, I have taken a small subset of the given dataset and check it for validity.**
- **For further analysis, I have displayed all the attributes and calculated the following parameters for each one of them:**

1. **Minimum Value of each attribute**
2. **1st Quartile for each attribute**
3. **Median Value for each attribute**
4. **Mean Value for each attribute**
5. **3rd Quartile for each attribute**
6. **Maximum Value of each attribute**